Lecture 10

Multiple Linear Regression & ANOVA

Statistics for
Civil & Environmental Engineers

Multiple Linear Regression(1)

- **Definition and Properties**
  - **A multiple linear regression model that relates a random response *Y* to (*p*-1) explanatory variables $x_1, x_2, \ldots, x_{p-1}$ takes the form**

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \epsilon$$

where β_0 , $\boldsymbol{\beta_1}$, …, β_{p-1}: p parameters

ε: random error term

Multiple Linear Regression(2)

- **Linear Least Squares Solutions Using the Matrix Method**
  - **The observations x_{ij} in the ($n \times p$) matrix:**

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np-1} \end{bmatrix}$$

  - **The multiple regression model:**
  - **The vector of mean values *E*[Y] of Y:**
  - **The least squares solution to the unknown parameters**
  - **The vector of estimated mean values of Y & residuals**

Multiple Linear Regression(3)

- **Properties of the Least Squares Estimators and Error Variance**
  - **Expectation of the least squares estimators**

Multiple Linear Regression(4)

- **Covariance matrix C of the least squares estimators**

$$C = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T]$$

where $E[\hat{\beta}] = (X^T X)^{-1} X^T E[Y]$

$$E[\beta] = (X^T X)^{-1} X^T Y$$

$$C = E[(X^T X)^{-1} X^T (Y - E[Y])(Y - E[Y])^T X (X^T X)^{-1}]$$

Multiple Linear Regression(5)

- **The error variance σ^2**

$$SS_E = \sum_{i-1}^{n} (y_i - \hat{y}_i)^2$$

$$SS_E = (y - X\hat{\beta})^T (y - X\hat{\beta})$$

$$= y^T y - \hat{\beta}^T X^T y - y^T X\hat{\beta} + \hat{\beta}^T X^T X\hat{\beta}$$

An unbiased estimator:

Multiple Linear Regression Example(1)

Example 6.6. Multiple regression on stream basin characteristics. Table E.6.1 gives some characteristics of 20 stream basins in the Valtellina region of Northern Italy. Physical evidence supports the hypothesis that in this area mean annual runoff is related to the mean annual rainfall and also to the mean elevation of the basin. The statistical significance of these relationships is considered in subsequent subsections. To formulate the model, therefore, we can treat the mean annual runoff as the response variable Y and the mean annual rainfall and the mean elevation as explanatory variables X_1 and X_2 respectively. Thus

$$\mathbf{X} = \begin{bmatrix} 1 & 1350 & 2329 \\ 1 & 1621 & 1593 \\ \vdots & \vdots & \vdots \\ 1 & 1283 & 2206 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} 1654 \\ 1374 \\ \vdots \\ 1023 \end{bmatrix}.$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1350 & 1621 & \cdots & 1283 \\ 2329 & 1593 & \cdots & 2206 \end{bmatrix} \begin{bmatrix} 1 & 1350 & 2329 \\ 1 & 1621 & 1593 \\ \vdots & \vdots & \vdots \\ 1 & 1283 & 2206 \end{bmatrix}$$

$$= \begin{bmatrix} 20 & 29,596 & 33,724 \\ 29,596 & 45,361,666 & 48,105,718 \\ 33,724 & 48,105,718 & 65,828,584 \end{bmatrix} = (\text{say}) \begin{bmatrix} a & b & c \\ b & e & f \\ c & f & g \end{bmatrix}.$$

taking into account the symmetry of the square $(p \times p)$ matrix $\mathbf{X}^T\mathbf{X}$.

The determinant of this matrix is $d = aeg + 2bcf - af^2 - b^2g - c^2e$. If the determinant is zero, the matrix is called *singular* and it does not have an inverse. If $(\mathbf{X}^T\mathbf{X})$ is nonsingular, its inverse is the transpose of the matrix of which the elements are the signed cofactors divided by the determinant. The inverse is thus given by

Multiple Linear Regression Example(2)

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} (eg - f^2)/d & (cf - bg)/d & (bf - ce)/d \\ (cf - bg)/d & (ag - c^2)/d & (bc - af)/d \\ (bf - ce)/d & (bc - af)/d & (ae - b^2)/d \end{bmatrix}$$

$$= \begin{bmatrix} 3.112081304 & -0.001509626 & -0.000491126 \\ -0.001509626 & 0.000000830 & 0.000000167 \\ -0.000491126 & 0.000000167 & 0.000000145 \end{bmatrix}.$$

Also,

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1350 & 1621 & \cdots & 1283 \\ 2329 & 1593 & \cdots & 2206 \end{bmatrix} \begin{bmatrix} 1654 \\ 1374 \\ \vdots \\ 1023 \end{bmatrix} = \begin{bmatrix} 25,661 \\ 38,852.792 \\ 45,285,738 \end{bmatrix}.$$

Therefore,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{bmatrix} -1035.2 \\ 1.0664 \\ 0.4390 \end{bmatrix}.$$

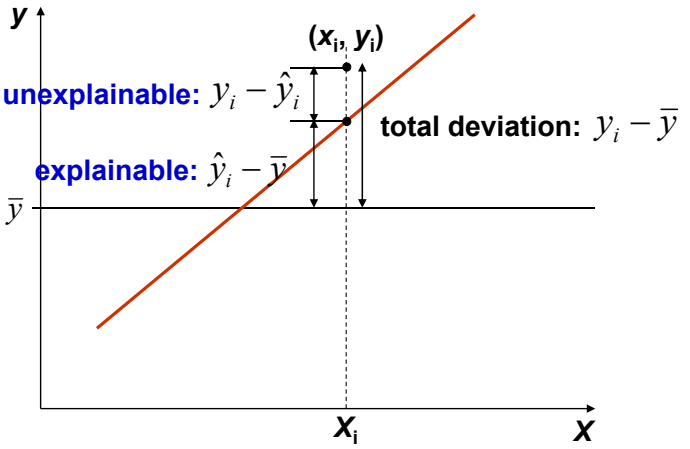
Hence the fitted model is $\hat{Y} = -1035.2 + 1.0664x_1 + 0.4390x_2$.

Note that roundoff errors will cause differences in solutions to Problems 6.6 through 6.16. Many software programs are available to find the alternate solutions accounting for roundoff errors and nonsingularities.[10]

Analysis of Variance (ANOVA)(1)

- **Properties**
  - To evaluate how good a regression model is, it is customary for statisticians to report an Analysis of Variance (ANOVA)
  - The analysis divides the observation variation in the data between that which can be explained by the regression relationship, and that which is residual unexplained error
- **Notation**
  - SS_T= total variation in the observed y values, called *total sum of squares*: $\sum_{i=1}^{n}(y_i-\bar{y})^2$
  - SS_R= *sum of squares* explained by the regression: $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$
  - SS_E= *sum of squares* attributed to the random error: $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

Analysis of Variance (ANOVA)(2)



Analysis of Variance (ANOVA)(3)

- **Relationship**

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2$$

Which can be written

SS_T = SS_R + SS_E

Total-SS = SS due to Regression + SS due to Error

With degree of freedom

n-1 = p-1 + n-p

Analysis of Variance (ANOVA)(4)

- **ANOVA Table**

Source of variation	Degree of freedom	Sum of squares	Mean square	*F* value
Treatment	p – 1	SS_R	$MS_R=SS_R/(p-1)$	MS_R/MS_E
Error	n – p	SS_E	$MS_E=SS_EE/(n-p)$	
Total	n – 1	SS_T		

Analysis of Variance (ANOVA)(5)

- **ANOVA variables**

ANOVA Example(1-1)

Q: Can we predict a student's weight *y* from his or her height *x*?

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
60	84	-8	-56	64	3136	448
62	95	-6	-45	36	2025	270
64	140	-4	0	16	0	0
66	155	-2	15	4	225	-30
68	119	0	-21	0	441	0
70	175	2	35	4	1225	70
72	145	4	5	16	25	20
74	197	6	57	36	3249	342
76	150	8	10	64	100	80
sum=612	1260			SS_{xx}=240	SS_{yy}=10426	SS_{xy}=1200

$\bar{x}$ =68 $\bar{y}$ =140

ANOVA Example(1-2)

The regression equation is: Weight= -200 + 5.00 Height

Predictor	Coef	Stdev	*t*-ratio	*p*
Constant	-200.0	110.7	-1.81	0.114
Height	5.000	1.623	3.08	0.018

ANOVA Table

Source	DF	SS	MS	*F*	*p*
Regression	1	6000.0	6000.0	9.49	0.018
Error	7	4426.0	623.3		
Total	8	10426.0			

SS_R SS_E SS_T

One-Way ANOVA(1)

- **One-Way ANOVA model**

$$X_{ij} = \mu + \theta_i + \varepsilon_{ij} \qquad i = 1,2,...,k \qquad j = 1,2,...,n_i$$

where μ: the overall mean, θ_i: the ith treatment effect
ε_{ij}: independent and normally distributed random error with zero mean and finite variance σ^2

- **Hypothesis**

$$H_0 : \theta_i = 0 \quad \text{for} \quad i = 1,2,...,k$$

$$H_1 : \theta_i \neq 0 \quad \text{for at least one } i$$

One-Way ANOVA(2)

- **Total Variability**

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2 \qquad \text{where} \qquad \bar{x}=\frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ij}$$

  - If H_0 is true, this variability is totally attributed to chance effects. Under H_1, on the contrary, a part of the variability can be attributed to differences between the treatment effects, which we estimate by

One-Way ANOVA(3)

- The test procedure applicable to Hypothesis is called ***analysis of variance***, because we partition the total variability in the data into different parts. The division is as follows:

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2=\sum_{i=1}^{k}\sum_{j=1}^{n}[(\bar{x}_i-\bar{x})+(x_{ij}-\bar{x}_i)]^2$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{x}_i-\bar{x})^2+2\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{x}_i-\bar{x})(x_{ij}-\bar{x}_i)$$

$$+\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2=n\sum_{i=1}^{k}(\bar{x}_i-\bar{x})^2+\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2$$

→ $SS_T=SS_{Tr}+SS_E$

One-Way ANOVA Table

TABLE 5.7.1
Analysis of variance for a one-way classification

Source of variation	Degrees of freedom	Sum of squares	Mean square	F value
Treatment	$k-1$	SS_{Tr}	$MS_{Tr}=\frac{SS_{Tr}}{(k-1)}$	$\frac{MS_{Tr}}{MS_E}$
Error	$k(n-1)$	SS_E	$MS_E=\frac{SS_E}{k(n-1)}$	
Total	$kn-1$	SS_T		

ANOVA Example(2-1)

Example 5.33. Compressive strength and density of concrete. We refer to the densities and compressive strengths of concrete listed in Table E.1.2. Our objective is to test whether compressive strength is a function of density. Suppose we treat density d as a single factor affecting the compressive strength and divide the densities into five levels or treatments as follows:

1. $d < 2430$ kg/m^3
2. $2430 \le d \le 2440$ kg/m^3
3. $2440 \le d < 2450$ kg/m^3
4. $2450 \le d < 2460$ kg/m^3
5. $d \ge 2460$ kg/m^3

Five "replicates" or observations are chosen for each treatment. Because experiments performed during a particular week (or longer period) may have some undesirable influences (such as operator bias) on the results, we shall try to spread the dates of observation as far as possible in the choice of observations from the set of 40 items. As seen, however, from Table 5.7.2 this effort is only partially successful in the case of the fifth treatment. Of course, this problem will not arise under a controlled experiment. We can in such a case randomize the effects of the experimenter, machinery, temperature, and other causal factors such as the moisture content of the concrete mix.

ANOVA Example(2-2)

TABLE 5.7.2
Table of concrete densities d, dates t, and strengths s from Table E.1.2

Treatment: density d, date t, strength s		Observations					Total	Mean
1	$d < 2430$ kg/m^3	2411	2415	2429	2428	2425		
	t	8/7	13/9	3/12	31/3*	26/6		
	s	58.8	50.7	68.1	56.9	59.8	294.3	58.86
2	$2430 \le d < 2440$ kg/m^3	2436	2436	2433	2436	2437		
	t	6/9	9/10	4/12	7/2*	21/9*		
	s	52.5	59.6	60.5	49.9	60.5	283.0	56.60
3	$2440 \le d < 2450$ kg/m^3	2445	2447	2445	2446	2448		
	t	9/9	23/9	14/10	18/12	19/3*		
	s	63.3	55.8	60.5	60.9	67.3	307.8	61.56
4	$2450 \le d < 2460$ kg/m^3	2454	2455	2454	2456	2456		
	t	12/7	3/9	3/10	6/12	9/3*		
	s	58.9	56.3	59.8	67.2	68.9	311.1	62.22
5	$d \ge 2460$ kg/m^3	2488	2473	2469	2472	2471		
	t	23/8	29/8	3/9	18/10	22/10		
	s	69.5	64.9	54.6	61.5	65.7	316.2	63.24
Total T and mean of s							1512.4	60.5

* Denotes that the year of testing is 1992; the other data pertain to 1991 tests.

ANOVA Example(2-3)

Null hypothesis H_0: The mean compressive strengths of concrete are the same.

Alternate hypothesis H_1: There are differences in the means.

We note that the two hypotheses are formally stipulated by Eqs. (5.7.1) and (5.7.2).

Level of significance $\alpha = 0.05$.

Calculations: The calculations are based on Tables 5.7.1 and 5.7.2 and Eqs. (5.7.5*a*) and (5.7.5*b*). Thus

$$SS_T = \sum_{i=1}^{5}\sum_{j=1}^{5} x_{ij}^2 - \frac{1}{25}T^2 = 92{,}217.14 - \frac{1512.4^2}{25} = 722.99.$$

Also,

$$SS_{Tr} = \frac{1}{5}\sum_{j=1}^{5} T_i^2 - \frac{1}{25}T^2 = \frac{458{,}207.98}{5} - \frac{1512.4^2}{25} = 147.45.$$

Hence

$$SS_E = 722.99 - 147.45 = 575.54.$$

The results are summarized in Table 5.7.3.

ANOVA Example(2-4)

TABLE 5.7.3
Analysis of variance for compressive strengths of concrete as a function of density

Source of variation	Degrees of freedom	Sum of squares	Mean square	F value
Density of concrete	4	147.45	$\frac{147.45}{4} = 36.86$	$\frac{36.84}{28.78} = 1.28$
Error	20	575.54	$\frac{575.54}{20} = 28.78$	
Total	24	722.99		

Decision: The F value is 1.28, which is less than $F_{4,20,0.05} = 2.87$ [where the degrees of freedom are $m = 5 - 1 = 4$ and $n = 5(5 - 1) = 20$]. Therefore, the null hypothesis is not rejected.

Coefficient of Determination

- **Definition**
  - **This is the ratio of the sum of squares due to regression to the total sum of squares**
  - **Sometimes called the *coefficient of multiple correlation***
  - **The statistic**